Acções auto-derrotadoras*

António Zilhão♣

1. Utilitarismo, maximização global, fraqueza de vontade, pré-compromisso e peculiaridade da condição humana

A função do coração é bombear sangue de forma eficiente para o sistema circulatório de um organismo cordato. Um coração que não o faça é um coração defeituoso, exactamente do mesmo modo que uma bomba de água que não bombeie água eficientemente para a canalização a que está ligada é uma bomba de água defeituosa. De modo semelhante, de um ponto de vista naturalista, uma mente defeituosa é uma mente que não realiza a sua função própria de modo adequado. Mas qual é a função própria da mente humana? E qual é o modo humano específico de realizá-la?

A resposta típica à primeira destas perguntas, sendo verdadeira, é todavia trivial. Trata-se da resposta de acordo com a qual a função própria da mente humana consiste em monitorizar tanto o ambiente externo como o metabolismo interno, por forma a guiar de forma eficiente as interacções que constantemente decorrem entre o organismo humano e o ambiente, sendo a finalidade última dessas interacções a sobrevivência, bem-estar e reprodução do organismo portador da mente. Para além de trivial, esta resposta à primeira pergunta tão-pouco nos fornece qualquer pista acerca de como responder à segunda.

Com efeito, o número de modos possíveis por meio dos quais pode ser desempenhada a função própria da mente humana, tal como ela é enunciada na resposta à primeira pergunta, é simplesmente gigantesco. Qual deles é o que dá efectivamente origem aos padrões específicos de comportamento que

* Este ensaio recolhe o texto de uma conferência que proferi no departamento de Filosofia da Universidade Nacional de Brasília (UnB) no dia 9 de Abril de 2013.

♣ Professor do Departamento de Filosofia da Faculdade de Letras da Universidade de Lisboa.

reconhecemos como 'humanos'? Se o nosso conhecimento da engenharia da máquina mental, isto é, do cérebro, se encontrasse num estádio bem mais avançado do que aquele em que de facto se encontra, poderíamos talvez usá-lo para delimitar e conduzir a investigação tendente a encontrar uma resposta para esta pergunta. Mas, no estádio actual de conhecimento neurofisiológico, isso não parece ser possível. Neste sentido, pelo menos por agora, a abordagem analítica continua a ser uma das ferramentas mais importantes a que podemos lançar mão na empresa de procurar compreender a mente humana e os comportamentos por meio dos quais ela se manifesta.

Podemos encontrar um dos desenvolvimentos mais completos desta abordagem no pensamento utilitarista; mais em particular, numa das perspectivas a ele associadas – a da chamada 'escolha racional'. Para além de outros, esta escola de pensamento tem um mérito inegável: providencia-nos uma caracterização, simultaneamente geral e, *prima facie*, plausível, por meio da qual parece ser possível capturar de fora, por assim dizer, o padrão que distingue os comportamentos humanos e a mente que os origina dos dos outros animais. Esta caracterização constrói-se em torno de dois elementos centrais: a procura da maximização global, por um lado, e a capacidade para o planeamento estratégico, por outro lado. Por outras palavras, a mente humana seria, desta perspectiva, nada menos do que uma máquina de maximização global realizada no cérebro humano pelo processo evolucionário. Um representante particularmente saliente desta tradição é Jon Elster, Professor na Universidade de Columbia, em Nova Iorque, e detentor, até 2011, da cátedra de Racionalidade e Ciências Sociais no Collège de France, em Paris.

O pensamento de Elster tem, todavia, uma particularidade que o torna especialmente interessante, pelo menos aos meus olhos. É que, não deixando de ser utilitarista, ele afasta-se do Utilitarismo tradicional num aspecto muito relevante. Este aspecto é o seguinte. De acordo com Elster, os dois traços definitórios da mente humana acima mencionados encontrar-se-iam apenas *imperfeitamente* realizados nos seres humanos efectivos. Esta imperfeição revelar-se-ia, mais especificamente, no facto de que os seres humanos sofreriam de uma insuficiência peculiar, a saber, a de, por vezes, incorrerem em *fraqueza de vontade*. Este conceito, por sua vez,

referiria um aspecto da acção humana que, à luz do ponto de vista utilitarista tradicional, teria que aparecer como surpreendente: o de que, por vezes, parece ser o caso que os homens agem contra as suas próprias preferências, tendo, além disso, consciência de que estariam a fazê-lo.

Por sua vez, a existência intuitiva de casos de acção humana com este contorno torna plausível a hipótese de que a nossa mente seria propensa a, sob certas circunstâncias, produzir acções auto-derrotadoras. Acontece que, de acordo com os já mencionados princípios fundamentais da perspectiva utilitarista, a hipótese de que uma tal propensão existiria nos seres humanos assume um carácter paradoxal. Esta é, sem dúvida, a razão pela qual muitos filósofos desta persuasão se esforçam por negar que a fraqueza de vontade seja mais do que apenas aparente. Não é o caso de Elster. Pelo contrário, ele distingue-a como um traço essencial da mente humana – trata-se, precisamente, daquela limitação que impediria os seus portadores de serem maximizadores globais perfeitos e que, por via disso, conferiria às suas acções uma estranha singularidade.

Esta singularidade resultaria, por sua vez, do facto de que, além de sofrer de fraqueza de vontade, o Homem saberia que esse é o caso. Deste modo, estando ciente da sua própria imperfeição, ele teria desenvolvido meios indirectos que lhe permitiriam ultrapassar a sua inconsistência primária e alcançar por caminhos ínvios os mesmos resultados que um maximizador global perfeito teria alcançado directamente. Seria então o recurso a estes meios indirectos que introduziria a originalidade peculiar da condição humana. Em simultâneo, ela permitiria também a Elster dissolver o paradoxo acima mencionado, ao permitir-lhe mostrar como a existência de acções humanas auto-derrotadoras poderia ser conciliada com o princípio geral de que os seres humanos se comportariam como maximizadores globais. De entre os meios indirectos desenvolvidos pela espécie humana para vencer a ameaça da inconsistência, os de maior relevo seriam, segundo Elster, as estratégias de pré-compromisso. Não é por isso de admirar que elas constituam o objecto central de estudo da sua teoria da racionalidade imperfeita. De acordo com esta teoria, por sua vez, um agente ou grupo de agentes pôr de pé uma estratégia de pré-compromisso consistiria, basicamente, em ele ou eles desencadearem um processo causal no mundo externo, o qual teria

como finalidade modificar de um modo pré-determinado o comportamento do seu próprio originador ou originadores.

Tal como usado por Elster, 'fraqueza de vontade' é, portanto, um termo bastante geral, cujo significado é integralmente definido pela caracterização introduzida acima; em particular, o uso deste termo não se encontra necessariamente ligado à descrição de alguma espécie de condição afligindo um órgão mental específico chamado de 'a vontade'. O próprio Elster fornece um rol de algumas das mais importantes manifestações comportamentais que cairiam sob a alçada deste termo, concebido de acordo com esta definição. Uma parte delas pertence ao catálogo dos fenómenos já diagnosticados na literatura clássica, nomeadamente, ser conduzido pela paixão, sentimentos ou apetite, em vez de pela razão; outra parte é de âmbito mais estratégico e deixa-se resumir por meio da ideia geral da sobreposição míope de um interesse próximo a interesses de mais longo alcance, mediados ou não por considerações de carácter moral, apesar do maior valor subjectivo destes últimos. Convém talvez introduzir aqui algumas considerações gerais para que possa tornar-se clara a importância deste segundo género de manifestações da fraqueza de vontade no âmbito da discussão que irá seguir-se.

Aquilo que é para ser maximizado por um maximizador global pode receber diferentes nomes. De acordo com a tradição à qual o pensamento de Elster pertence, o mais comummente usado é 'utilidade'. Assim, no interior desta tradição, o modo habitual de entender um maximizador global é como alguém que tenta maximizar a utilidade que atribui aos desfechos esperados das suas acções em qualquer momento do tempo. Acontece, porém, que a vida humana se desenvolve ao longo de um contínuo de pontos temporais. Assim sendo, e como Elster correctamente salienta, para que esta ideia tradicional possa ter a pretensão de ser útil de um ponto de vista teórico, ela necessita de ser ampliada de modo a que passe a incluir considerações temporais no seu seio. Por sua vez, a introdução da temporalidade no processo analítico traz consigo a descoberta de novas formas peculiares por meio das quais um agente pode falhar a maximização da utilidade. Mais uma vez, é o próprio Elster quem passa em revista algumas delas. Há três que merecem aqui ser destacadas: as mudanças de preferência sem motivo, a inconsistência temporal estratégica e o desconto

hiperbólico.

Do ponto de vista utilitarista, estas formas de falhar a maximização da utilidade não podem também deixar de cair sob a definição acima apresentada de fraqueza de vontade. Deste modo, e de forma não surpreendente, Elster defende que o Homem lida com elas através do mesmo género de processos indirectos já descrito acima, isto é, limitando-se a si mesmo por meio da adopção de estratégias de pré-compromisso. Elster escolheu mesmo um símbolo literário para personificar esta forma peculiarmente humana de alcançar a racionalidade. Trata-se do Ulisses do canto XII da Odisseia, o qual ordena aos seus companheiros que o atem fortemente ao mastro da embarcação e, simultaneamente, tapem os seus próprios ouvidos com cera, de tal modo que ele possa escutar o canto das sereias, sobreviver à experiência, e manter constante o rumo do seu barco.

2. Utilitarismo, desconto exponencial, fraqueza de vontade, desconto hiperbólico e picoeconomia

Como disse acima, o interesse principal de Elster no desenvolvimento da sua teoria da racionalidade imperfeita é o de estudar as diferentes formas de pré-compromisso que podem ser observadas no comportamento humano, tanto individual como socialmente. Deste modo, ele está interessado não tanto em desenvolver uma teoria da fraqueza de vontade quanto em desenvolver uma teoria das estratégias humanas para lidar com ela.
Todavia, dada a extrema importância que o fenómeno da fraqueza de vontade adquire, seja na abordagem de Elster, seja no contexto do pensamento utilitarista em geral, ele merece ser melhor compreendido nos seus próprios termos.

Uma excelente tentativa de produzir uma abordagem teórica simultaneamente abrangente e consistente acerca deste fenómeno pode ser encontrada no trabalho de George Ainslie, professor de Psiquiatria clínica na Temple University, em Philadelphia. Vamos então dar uma olhadela ao que ele tem para nos dizer acerca deste tópico.

A tese original de Ainslie é a de que todas as razões para o pré-compromisso passadas em revista por Elster, isto é, todas as

formas de fraqueza de vontade por ele descritas, admitem ser, em última análise, reconduzidas a um único problema, mais profundo do ponto de vista explicativo, a saber, o problema do desconto hiperbólico. Note-se que, tal como referi acima, Elster também trata do desconto hiperbólico no seu próprio trabalho, mas fá-lo no âmbito de uma descrição de uma forma específica de fraqueza de vontade, e não com vista a reconduzir este conceito àquele. De acordo com Ainslie, porém, e com bons argumentos, a noção de fraqueza de vontade seria apenas uma outra forma, mais intuitiva, de conceptualizar o fenómeno do desconto hiperbólico. Mas de que fala um teórico da escolha racional quando fala de 'desconto hiperbólico'?

A primeira coisa a dizer a este respeito é a de que a noção de desconto hiperbólico deve ser considerada por contraste com a noção de desconto exponencial. Esta, por sua vez, refere o modo como um maximizador global deve, em teoria, descontar o futuro. Permitam-me, então, que, para responder à pergunta acima, comece por clarificar um pouco melhor esta noção de desconto exponencial do futuro.

Como todos sabemos, enquanto que o presente é certo, o futuro é incerto. Infelizmente, é sempre possível que um agente hoje vivo e de boa saúde deixe de estar vivo num futuro não muito longínquo ou até mesmo num futuro próximo. Ora, agentes mortos não têm a oportunidade de maximizar o que quer que seja. Obviamente, se o futuro contemplado for suficientemente distante, é também certo para um qualquer agente que ele já não estará vivo quando esse futuro se tornar presente. Mas, entre estas duas certezas, há um largo campo de incerteza.

Faz, por isso, todo o sentido que um maximizador global desvalorize satisfações diferidas de utilidade. Todavia, dado que, na ausência de uma morte inesperada, os futuros passados se tornam, com o passar do tempo, presentes vivos, e o presente vivo se torna num passado morto, recordado apenas com maior ou menor nitidez, para que o comportamento de um maximizador global se mantenha consistente ao longo do tempo, o modo como essa desvalorização se processa tem que obedecer a alguma espécie de ordem. E a ordem que preserva a consistência e, portanto, a racionalidade, é, precisamente, aquela que se deixa representar por meio do uso de uma função exponencial da taxa de desconto que o agente emprega

na desvalorização da utilidade futura.

Deixem-me introduzir aqui um exemplo que, espero, vos ajudará a compreender por que é que a abordagem racional do futuro deve ser feita em termos da ideia de desconto exponencial[1]. Imaginem que ler este ensaio perante vós vale, digamos, 100 unidades de utilidade para mim neste preciso momento. Imaginem também que a minha taxa de desconto para a leitura de ensaios em palestras é de 10% por semana. Sob cada um destes pressupostos, a perspectiva de ler este ensaio perante vós hoje teria valido 90 unidades de utilidade há uma semana atrás, 81 unidades de utilidade há duas semanas atrás, 72,9 unidades de utilidade há três semanas atrás, e assim sucessivamente.

Estes valores são obtidos pela computação de uma função exponencial da taxa de desconto, como a indicada abaixo[2]. É precisamente por isso que esta função é chamada de 'exponencial'. Suponham agora que ler ensaios perante uma audiência internacional tem também um custo para mim, o qual necessita de ser pago na semana subsequente à conferência sob a forma de inveja, ressentimento e retaliação administrativa perpetradas pelos meus colegas no meu departamento de origem. Suponhamos que esse custo é de 110 unidades de utilidade. Se supusermos que este é, de facto, o caso, e que a taxa de desconto se mantém a mesma, então o valor liquido de me dirigir a vós, hoje, é de 100-99 unidades de utilidade, isto é, é de 1 unidade de utilidade (i.e., de {100-[110.(1-0.1)]}=[100-(110.0.9)]=(100-99)). Portanto, hoje, eu deveria ter vindo e apresentado este ensaio perante vós.

Suponhamos agora que eu contemplei esta perspectiva na semana passada. O valor líquido de estar aqui hoje teria sido de 90-89.1 unidades de utilidade , i.e., de 0.9 unidades de utilidade (i.e., de [(100.0.9)-(110.0.9.0.9)] = 90-89.1=0.9). Logo, há uma semana atrás eu deveria ter decidido vir hoje. E se supusermos que eu considerei a perspectiva de estar aqui hoje há duas semanas atrás, o valor liquido de estar aqui hoje teria então sido de 81-80.19 unidades de utilidade, i.e., teria sido de 0.81 unidades de utilidade

[1] Este exemplo adapta ao caso que aqui introduzo um exemplo apresentado pelo próprio Ainslie.

[2] A fórmula que exprime esta função particular é a seguinte:

$$\text{Valor}=\text{Valor Objectivo}.\,(1-\mathit{TaxadeDesconto})^{\mathit{Diferimento}}$$

(i.e., de [(100.0.9.0.9)-(110.0.9.0.9.0.9)]=81-80.19=0.81). Portanto, com o desconto exponencial, a diferença entre a utilidade de me dirigir a esta audiência hoje e os custos em que incorrerei na semana subsequente por o ter feito vai diminuindo progressivamente à medida que aumenta a antecedência com a qual a perspectiva de o fazer é contemplada. O que é crucial, todavia, é que ela nunca se torna negativa ou chega sequer a zero. Assim, se, hoje, eu escolhi apresentar esta palestra, a despeito das consequências futuras de tê-lo feito, então eu teria também escolhido apresentá-la há uma, duas, ou três semanas atrás ou, mesmo, há dois meses atrás, como de facto aconteceu. E se hoje tivesse escolhido não vir – se, por exemplo, o custo das consequências futuras valesse, para mim, 120 unidades de utilidade em vez das 110 indicadas acima – então eu teria igualmente escolhido a uma qualquer distância temporal do presente não estar aqui hoje. Em resumo, só se a utilidade futura for descontada por meio de uma função exponencial da taxa de desconto usada para desvalorizá-la é que a consistência ao longo do tempo fica garantida; e só poderemos ser maximizadores globais se formos consistentes ao longo do tempo.

Esclarecida a noção de desconto exponencial, e por que é que é ela que constitui a forma que uma taxa de desconto do futuro deve assumir para ser racional, vou então passar agora a apresentar a noção de desconto hiperbólico. Uma função de desconto hiperbólica é uma função de desconto que se deixa representar graficamente por uma curva muito mais arqueada do que a que representa uma função de desconto exponencial. Este maior arqueamento significa que, enquanto que, no caso tanto de diferimentos muito longos como de diferimentos muito curtos, não haverá qualquer diferença no valor atribuído por ambas as funções aos mesmos desfechos, esse valor será, todavia, largamente dissemelhante nos pontos intermédios. Na realidade, de acordo com a função de desconto hiperbólica, os desfechos intermédios serão menos valorizados do que o serão de acordo com a função de desconto exponencial.[3]

[3] Uma fórmula hiperbólica de acordo com a qual o valor descontado quando não há qualquer diferimento é o mesmo que o valor objectivo é, por exemplo, a seguinte: Valor= $\frac{ValorObjectivo}{(1+Diferimento)}$. O valor descontado de um desfecho a

Que tem então a ideia de desconto hiperbólico a ver com a ideia de fraqueza de vontade? A conexão é a seguinte. Segundo Ainslie, inúmeras observações empíricas por ele efectuadas do comportamento de agentes humanos em situações reais contradisseram claramente o pressuposto de que as preferências temporais espontâneas de um agente típico seriam enquadráveis num qualquer tipo de representação exponencial. Foi então a tentativa de encontrar um enquadramento teórico alternativo, por meio do qual ele pudesse dar conta dessas observações empíricas, que levou Ainslie à descoberta de que essas preferências poderiam ser representadas por meio de uma função de desconto hiperbólica. Ora, a opção pelo uso de uma tal função na modelação da dimensão temporal do comportamento de escolha dos agentes por ele estudados acarretou duas consequências de grande relevo para a definição da sua racionalidade.

A primeira destas consequências é a seguinte: imaginemos um agente que desconta os desfechos prospectivos das suas acções hiperbolicamente, e consideremo-lo independentemente de qualquer interacção com outros agentes. Se lhe for pedido que escolha entre obter um desfecho de menor valor com algum diferimento D ou obter um desfecho do mesmo género, mas de maior valor, que só estará disponível mais tarde com um diferimento D', mas não ambos, então ele começará por escolher o desfecho de maior valor quando D é longo, mas reverterá para o desfecho de menor valor quando D ficar suficientemente curto. Assim, ele deixará de poder obter o que, intuitivamente, parece ser o desfecho óptimo, a saber, o desfecho de maior valor com um diferimento D', apesar de ter escolhido inicialmente proceder dessa forma.

A segunda das consequências acima mencionadas é a seguinte: imaginemos o mesmo agente a interagir com outros agentes e, em particular, com agentes que descontam o futuro de acordo com uma função de desconto exponencial. Suponhamos que estes agentes se envolvem nalguma espécie de comércio. Dado que, como foi dito acima, nos diferimentos muito longos ou muito curtos não haverá qualquer diferença no valor associado aos mesmos

uma unidade de diferimento seria de 50% do seu valor objectivo, a duas unidades de diferimento seria de 33% do seu valor objectivo, a três unidades de diferimento seria de 25% e assim sucessivamente.

desfechos quando considerados por ambas as funções, mas que os desfechos intermédios serão avaliados como tendo menor valor de acordo com a função de desconto hiperbólica, então o agente que desconta o futuro exponencialmente pode abusar sistematicamente do agente que desconta o futuro hiperbolicamente comprando-lhe por um preço barato os bens de que ele dispõe a uma distância temporal suficientemente afastada do seu uso prospectivo e revendendo-lhos a um preço muito mais alto a uma distância temporal suficientemente próxima da ocasião de uso dos bens em questão.

Como é bom de ver, tanto um agente que, pela mecânica do seu próprio processo de escolha, é incapaz de escolher de entre duas opções a que tem um maior valor para ele, como um agente que tem uma predisposição para que os seus recursos sejam bombeados da sua posse por um processo como o descrito acima, só muito dificilmente poderão merecer ser classificados como maximizadores globais, mesmo que de um tipo apenas imperfeito. Na realidade, o uso de uma função hiperbólica para representar o modo como estes agentes descontam o futuro mostrou-se ser uma forma de representar um padrão de comportamento que constitui uma instância de fraqueza de vontade, de acordo com a definição deste termo apresentada acima.

Ciente destas consequências e da sua importância, Ainslie defende então que a análise dos resultados empíricos por ele obtidos mostra que o Homem é, na realidade, uma criatura que, com frequência, adopta padrões auto-derrotadores de comportamento e, que, mais ainda, está condenado pela própria Natureza a agir deste modo. Para alguém oriundo da tradição intelectual do Utilitarismo, como é o seu caso, a defesa deste género de posições não pode deixar de vir acompanhada da extracção de importantes consequências teóricas.

Como vimos, um dos pressupostos nos quais a teoria utilitarista tradicional se baseia é o pressuposto de acordo com o qual o Homem é essencialmente um maximizador global, consistente ao longo do tempo. Tipicamente, a aceitação deste pressuposto não deixa ao pensamento utilitarista outra opção para dar conta das falhas de racionalidade efectivamente observadas nos agentes empíricos que não seja a de apelar para explicações das mesmas em termos de lapsos de computação ou outras formas de

mau funcionamento episódico do aparelho cognitivo. Mas um tal modo de dar conta de um fenómeno tão difundido e regular quanto o é a fraqueza de vontade é claramente insatisfatório, se não mesmo contra-intuitivo, como o próprio Ainslie foi levado a reconhecer, pressionado pelas suas próprias investigações em Psicologia empírica.

Deste modo, seguindo na esteira de Elster, mas desenvolvendo numa nova direcção a sua intuição original acerca da natureza do carácter indirecto da racionalidade humana, Ainslie propõe uma reforma de fundo do Utilitarismo. Trata-se da sugestão de que o pressuposto de que a acção humana seria naturalmente regida por regras de consistência inter-temporal deveria ser deixado cair. Esta é, sem dúvida, uma reforma de monta, na medida em que implica o abandono definitivo de um dos pressupostos com base nos quais se define explicitamente a perspectiva de que o autor se reivindica. Consistente com ela, Ainslie defende que o pensamento utilitarista deveria passar a ser entendido como baseando-se apenas num pressuposto – o de que a mola mental do comportamento e da acção humanos seria a motivação e não o juízo.

Uma consequência da reforma do Utilitarismo proposta por Ainslie é a de que o aspecto problemático que fica agora por descrever é, não o modo particular, indirecto, como os seres humanos implementariam a sua natureza de maximizadores globais, a qual seria, segundo ele, fundamentalmente inexistente, mas antes o modo como criaturas cujas estruturas mentais descontariam instintivamente o futuro de forma hiperbólica, isto é, criaturas que, do ponto de vista da racionalidade temporal, seriam naturalmente irracionais, poderiam dar a impressão de comportar-se frequentemente como se fossem verdadeiros descontadores exponenciais, isto é, como se fossem temporalmente racionais. Por outras palavras, e em oposição ao título do ensaio clássico de Donald Davidson, o problema a necessitar de elucidação teórica teria deixado de ser o de descobrir quais seriam as condições de possibilidade para a existência de fraqueza de vontade em maximizadores globais, para passar a ser o de determinar os processos por meio dos quais seria possível que criaturas cujas estruturas mentais seriam naturalmente incontinentes produzissem a ilusão de que se comportariam como se fossem criaturas continentes. E é precisamente isto que a sua própria teoria da

racionalidade indirecta se propõe fazer. Num certo sentido, esta teoria de Ainslie pode ser vista como uma variação sobre o tema freudiano de tentar reconstruir o Utilitarismo de um modo tal que ele passe a ser capaz de dar conta de um modo plausível da irracionalidade, entendida como uma dimensão essencial da condição humana.

Ao contrário de Elster, porém, que enfatiza fortemente a importância da criação de estruturas externas para limitar os agentes humanos e constrangê-los a comportarem-se como se fossem directamente racionais, Ainslie defende que esta característica aparente do comportamento humano admite ser integralmente entendida a partir de dentro do próprio plano mental individual. Tal como no caso de Freud, é a subdivisão da mente em múltiplos agentes autónomos que vai proporcionar-lhe a solução para o seu problema. Com efeito, ele defende que a aparência frequente de racionalidade temporal directa no comportamento dos seres humanos não seria senão o resultado indirecto de uma mistura de conflito e negociação que ocorreria continuamente no interior da mente humana entre interesses particulares autónomos e rivais que viveriam juntos nela e que, nela, estariam constantemente a tentar derrotar-se uns aos outros e a tentar preponderar. Por outras palavras, segundo Ainslie, quando presente, a racionalidade humana indirecta resultaria, no plano individual, da mesma espécie de mecanismo de mão invisível que, de acordo com Adam Smith, originaria a racionalidade social de uma economia de mercado. Em consequência, a mente humana exibiria precisamente o mesmo tipo de êxitos e colapsos que é habitual os mercados exibirem. Esta é, aliás, a razão pela qual ele chama à sua abordagem da mente humana '*picoeconómica*'.

3. Encontro de probabilidades: um exemplo de comportamentos e acções auto-derrotadoras mas não dependentes de considerações relacionadas com a preferência temporal

As propostas de reforma do Utilitarismo que acabei de sumariar são bastante apelativas. Creio, todavia, que o modo como se propõem dar conta do problema da acção auto-derrotadora tem uma falha importante: restringe demasiado o âmbito de aplicação deste conceito. Isto acontece por forma a que a solução proposta

para o problema original seja aceitável do ponto de vista do único dos pressupostos do Utilitarismo tradicional que foi deixado de pé. Assim, na parte remanescente deste ensaio, vou tentar apresentar-vos um modo diferente e, creio, mais apropriado, de dar conta da origem mental da existência destes comportamentos e acções.

Começo esta apresentação pela descrição de um caso no qual me parece ser claro que o desempenho de um tipo particular de acção auto-derrotadora, por um lado, e a procura de gratificação imediata, ou próxima, por outro lado, nada têm a ver um com o outro. Em seguida, irei sugerir um modo de explicá-lo. Após o que defenderei que a explicação que apresento para ele também pode dar conta dos casos considerados por Elster e Ainslie. Finalmente, concluirei afirmando que o âmbito mais amplo da minha explicação a torna preferível às reconstruções da estrutura da mente e da acção humanas apresentadas tanto por Elster como por Ainslie.

Considerem então o meu caso. Vou chamar-lhe o caso do "encontro de probabilidades". Foi tornado saliente num experimento efectuado por Gallistel com estudantes de Psicologia em Yale. Vale a pena considerar este experimento atentamente pelo que vou expô-lo com algum detalhe. É então o seguinte.

Um rato foi treinado para percorrer um labirinto em forma de T; este labirinto dispunha de alimentadores montados nas extremidades de cada um dos seus braços. O alimentador do lado esquerdo era fornecido com comida em 75% dos ensaios; o alimentador do lado direito era fornecido com comida em 25% dos ensaios. A distribuição sequencial de comida nos alimentadores não obedecia a um padrão pré-determinado. Se o rato escolhesse o alimentador contendo comida, comê-la-ia; se escolhesse o alimentador vazio, não receberia qualquer comida. Sobre cada um dos alimentadores encontrava-se uma lâmpada eléctrica, colocada dentro de um resguardo; esta lâmpada acender-se-ia se o alimentador fosse fornecido com comida e manter-se-ia apagada se o alimentador permanecesse vazio. O rato não poderia ver a luz quando a lâmpada estivesse acesa, nem a sua ausência quando ela estivesse apagada; mas os estudantes que o observavam poderiam. Antes do início de cada ensaio, os experimentadores pediam aos estudantes que elaborassem uma previsão acerca de qual das lâmpadas se iria acender nesse ensaio.

O que o experimento mostrou foi que o rato rapidamente

adoptou a estratégia de escolher sempre o lado no qual a probabilidade de encontrar comida seria maior. Quer dizer, ele escolheu em quase 100% dos ensaios dirigir-se para o lado esquerdo do labirinto e em quase 0% dos ensaios dirigir-se para o lado direito do mesmo. Deste modo, o rato maximizou efectivamente a sua taxa de sucesso. Conseguiu ser alimentado em 75% dos ensaios.

Já os estudantes nunca adoptaram a estratégia do rato. Eles fizeram coincidir sistematicamente a frequência relativa das suas escolhas com as frequências relativas com as quais cada um dos lados era fornecido com comida. Quer dizer, eles escolheram o lado esquerdo em aproximadamente 75% dos ensaios e o lado direito em aproximadamente 25% dos ensaios. Dado que cada escolha do lado esquerdo tinha uma probabilidade de 0,75 de ter êxito e uma probabilidade de 0,25 de não o ter, e que cada escolha do lado direito tinha uma probabilidade de 0,25 de ter êxito e uma probabilidade de 0,75 de não o ter, a sua taxa de sucesso foi de 62,5%, consideravelmente abaixo da do rato. Os experimentadores expuseram em seguida aos estudantes o modo como a obtenção deste resultado "demonstrava" que o rato tinha exibido um comportamento mais inteligente do que eles. Os estudantes ficavam então livres para tirar a conclusão que se impunha …

Mas esta história não acaba aqui. De facto, os problemas com que o rato e os estudantes foram confrontados não eram exactamente iguais e os experimentadores sabiam-no. Com efeito, o rato não tinha qualquer contacto sensorial com a extremidade do braço do labirinto que não tinha escolhido em cada ensaio; portanto, não podia saber que, quando não havia comida na extremidade do braço que ele tinha escolhido, haveria comida na extremidade do outro braço. Mas os estudantes sabiam sempre que esse era o caso.

Espantosamente, quando o experimento foi subsequentemente modificado, de tal forma que o rato passasse a poder ter sempre contacto sensorial com a extremidade do braço que não tinha escolhido, ele rapidamente alterou o seu comportamento; tal como os estudantes, ele começou a fazer coincidir as frequências relativas das suas escolhas com as frequências relativas com as quais as extremidades de cada um dos braços eram fornecidas com comida.

A percepção de que, quando verdadeiramente colocado na mesma situação cognitiva na qual eles tinham sido colocados, o rato

exibia o mesmo padrão de escolha que eles, foi sem dúvida motivo de grande alívio e conforto para os estudantes. Mas, independentemente de ter espoletado esta bem-vinda sensação nos estudantes, a exibição deste novo padrão de comportamento por parte do rato levanta enormes perplexidades ao investigador. Duas questões, em particular, necessitam de ser respondidas: por que é que o rato deixou de maximizar a sua taxa de sucesso na segunda versão, apenas ligeiramente modificada, do experimento? E por que é que, tendo-o feito, ele substituiu a sua estratégia óptima original por exactamente a mesma estratégia subóptima que se descobriu que os estudantes seguiam espontaneamente? O mistério que envolve a colocação destas duas questões adensa-se ainda mais quando se constata que, de acordo com diversos etologistas, este tipo de comportamento é comummente observado em diferentes espécies animais.

Este é claramente um caso no qual a racionalidade dos agentes, compreendida em termos de maximização da utilidade individual, parece colapsar. Todavia, e ao contrário dos casos de fraqueza de vontade passados em revista por Elster e Ainslie, falhas de racionalidade deste género não admitem ser descritas como tendo sido causadas por uma incapacidade dos agentes de evitarem a procura de gratificação imediata ou, pelo menos, próxima, a despeito da presença neles de uma intenção para dirigirem os seus esforços no sentido da obtenção de um bem melhor, mas de aquisição mais diferida no tempo. Neste sentido, parece claro que o problema com que nos encontramos confrontados neste caso não é um problema de determinar como é que o valor de desfechos futuros deve ser descontado para o seu valor presente. Assim sendo, necessitamos de olhar noutra direcção para encontrar uma explicação para comportamentos e acções auto-derrotadoras deste género. Mas como fazê-lo?

Na realidade, o princípio que precisamos de seguir para começar a encontrar uma resposta a esta pergunta é bastante simples, embora nem sempre seja fácil de pô-lo em prática. Trata-se do princípio de que, para compreender efectivamente o *rationale* de um comportamento ou acção, é necessário dirigir a nossa atenção para fora do laboratório e focá-la na análise das situações naturais ou sociais reais nas quais ele ocorre. Se o fizermos neste caso, constataremos que, ao contrário do que aconteceu no caso do rato

de Yale, mantido sozinho no seu labirinto em forma de T, em circunstâncias naturais, um local onde um animal encontra comida com frequência, ou em grandes quantidades, é um local que, com toda a probabilidade, é também conhecido e frequentado por outros animais residentes na mesma área e alimentando-se do mesmo tipo de comida.

Ora, esta constatação faz toda a diferença. Com efeito, se todos os animais que se alimentam do tipo de comida em causa adoptarem a estratégia de escolherem sempre procurar comida no local onde a experiência anterior indica que esta se encontrará disponível com maior abundância ou frequência, negligenciando em simultâneo os locais onde sabem que a comida será menos abundante ou menos frequente, a consequência inevitável será a de que a quantidade de comida disponível para cada animal naquele local começará a diminuir numa proporção inversa ao do constante aumento do número de animais que convergem para ele em busca de comida. Em última instância, um animal que procure comida onde todos sabem que esta é menos frequente ou menos abundante, mas onde os competidores serão também em menor número ou aparecerão com menor frequência, acabará por gozar de uma vantagem adaptativa. Esta análise simples mostra assim claramente que seguir uma estratégia de maximizar a utilidade esperada numa situação na qual todos os indivíduos competem pelo mesmo tipo de recurso não é uma estratégia evolucionariamente estável. Na realidade, é possível mostrar por meios analíticos que, em situações como a acabada de descrever, a única estratégia evolucionariamente estável que se encontra à disposição dos agentes é, precisamente, a estratégia do encontro de probabilidades. Por outras palavras, o encontro de probabilidades revela-se ser a única estratégia, cuja adopção por todos os agentes não gera condições que criam um ambiente selectivo contra ela própria.

Tendo descoberto qual o *rationale* subjacente à adopção generalizada da estratégia do encontro de probabilidades no reino animal, encontramo-nos então agora em posição de conseguir perceber que a resposta dada pelos estudantes *e pelo rato* ao problema com que só os primeiros foram inicialmente confrontados, apesar de não ser apropriada na situação laboratorial peculiar inteligentemente imaginada pelos experimentadores, *é apropriada* em circunstâncias que, embora não sejam idênticas, são muito

semelhantes a ela, e que são, além disso, muito mais prováveis de ocorrer num ambiente natural. Sob *essas* condições, a sua resposta teria sido objectivamente racional.

Ora, o facto de esta resposta ter sido desempenhada espontaneamente e sem esforço pelos estudantes, por um lado, e o facto de se ter tratado de uma instância de um comportamento tão difundido no reino animal, por outro lado, faz-me pensar que estamos aqui perante um fenómeno peculiar. Que fenómeno será esse?

4. Uma resposta cognitivista de amplo espectro ao problema da acção auto-derrotadora

Antes de começar a responder à pergunta que acabei de formular, resposta esta que permitirá, por sua vez, concluir a resposta à pergunta anterior, deixem-me introduzir aqui algumas considerações acerca do modo como Ainslie vê o alcance do seu trabalho. Como disse acima, ele tem de si mesmo a imagem de um renovador do Utilitarismo. Enquanto tal, ele considera-se um contribuinte para o debate plurissecular que, desde o século XVIII, tem vindo a ser travado entre utilitaristas e cognitivistas. Mas, como vimos, dado que Ainslie abandonou o pressuposto da maximização global, ao qual Elster se mantém fiel, mesmo que de um modo muito peculiar, o Utilitarismo passa para ele a definir-se apenas como a perspectiva de acordo com a qual a experiência e a expectativa da recompensa são as únicas molas mentais essenciais ao desencadear da acção e do comportamento humanos. E o Cognitivismo torna-se, por isso, por contraste, a perspectiva de acordo com a qual essas molas são o juízo e a avaliação. Ao apresentar uma sugestão de resolução do problema da fraqueza de vontade, e da acção auto-derrotadora em geral, dentro dos limites de uma abordagem estritamente motivacional da explicação do comportamento, Ainslie clama então que conseguiu "levar a teoria da utilidade ao desempate no seu confronto com o Cognitivismo." (BoW, p.38).

Discordo dele. O exemplo do encontro de probabilidades que há pouco apresentei pretende, precisamente, mostrar como o domínio da acção auto-derrotadora não se deixa de todo esgotar pelo género de casos de fraqueza de vontade que Ainslie considera

serem os relevantes. Na realidade, muita da investigação empírica recente na área da chamada "psicologia da irracionalidade" tem trazido à luz de forma consistente um grande número de exemplos que poderiam ser adicionados ao que eu aqui trouxe à vossa consideração.

Assim, mesmo que reconheçamos a Ainslie o mérito de ter encontrado uma solução interessante para aqueles casos que, de facto, considerou, a sua abordagem ainda nos deixa às escuras acerca do modo como devemos estendê-la aos casos de comportamento auto-derrotadora nos quais o problema subjacente não cai de todo sob o figurino clássico do problema de o agente 'sucumbir à tentação' da gratificação próxima e, portanto, não admite ser explicado à custa de considerações de carácter temporal modeladas por meio de uma função de desconto hiperbólica. Não creio, por isso, que a teoria da utilidade tenha alcançado o tal desempate contra o Cognitivismo reivindicado por Ainslie. Pelo contrário, se aceitarmos os termos do debate tal como são apresentados por ele, creio mesmo que, em contradição com as suas palavras, o Cognitivismo já alcançou esse desempate há bastante tempo, como passarei a tentar demonstrar.

Antes de fazê-lo, preciso ainda de adiantar algumas palavras acerca do que devemos entender pelo termo 'Cognitivismo'. Isto é tanto mais necessário quanto as objecções de Ainslie à abordagem cognitivista deixam claro que o que ele entende por 'Cognitivismo' é um modelo clássico, já desactualizado, de racionalidade ilimitada, no contexto do qual um eu consciente, concebido basicamente como uma espécie de 'fantasma na máquina', para usar a expressão de Ryle, está constantemente a integrar dados motivacionais, avaliativos e cognitivos com vista a tentar computar uma qualquer solução óptima para o problema que tem entre mãos, antes de dar início à acção. Mas há outras abordagens cognitivistas para o problema da explicação do comportamento humano e animal que são bastante mais interessantes e realistas do que esta. De entre estas, as mais importantes são, do meu ponto de vista, aquelas que foram desenvolvidas do interior do ponto de vista da chamada 'racionalidade limitada'.

Deste ponto de vista, a abordagem que, por sua vez, me parece ser a mais promissora é a das chamadas 'heurísticas rápidas e frugais', desenvolvida por Gigerenzer e o seu grupo ABC no

Instituto Max Planck, em Berlim. Esta abordagem admite ser designada como 'cognitivista', no sentido em que defende que a acção é determinada por juízos do agente. Mas os mecanismos que subservem a formação desses juízos são totalmente dissemelhantes daqueles que são postulados pelas teorias cognitivistas tradicionais. De acordo com este ponto de vista, a mente humana é vista como estando organizada como se fosse uma 'caixa de ferramentas adaptativa', isto é, como consistindo num conjunto de heurísticas especificamente adaptadas para lidar com os problemas cognitivos que assumiram um significado fundamental no decurso do passado evolucionário do Homem. Por sua vez, estas heurísticas caracterizam-se por tirarem partido de capacidades evolvidas que se encontram naturalmente disponíveis nos seres humanos, com vista a encontrarem soluções simples, mas efectivas, para os problemas com que estes tiveram que lidar nos ambientes nos quais viveram.

Ora bem, como vimos, os teóricos do Utilitarismo explicam tipicamente a possibilidade da acção auto-derrotadora pela incapacidade dos agentes de evitarem procurar a gratificação iminente ou, pelo menos, próxima, a despeito da presença neles de uma intenção de procurar um bem melhor, isto é, mais gratificante, embora de obtenção mais distante no tempo. Dado que se supõe que o bem em questão é, mesmo se descontado ao seu valor presente, subjectivamente melhor que aquele que os agentes efectivamente escolheram, então, inevitavelmente, estes são casos nos quais, das duas, uma: ou os agentes não levaram em conta a sua própria escala de preferências; ou a máquina cognitiva dos agentes deixa de funcionar apropriadamente e falha na computação da maximização da utilidade. Como também já vimos, o problema associado a cada uma destas explicações está em que a primeira contradiz frontalmente um dos principais pressupostos da teoria supostamente explicativa e a segunda é inerentemente implausível, dada a larga difusão do fenómeno entre os humanos. Deste modo, sempre que têm que lidar com a questão da acção auto-derrotadora, os teóricos da interpretação racional encontram-se perante um nó górdio. É precisamente este nó górdio que tanto Jon Elster como George Ainslie tentaram desatar no seu trabalho por meio do seu apelo ao conceito de racionalidade indirecta.

Mas, ao focar-se apenas em considerações de racionalidade inter-temporal para enfrentar esta questão, a perspectiva da

racionalidade indirecta é impotente para dar conta daqueles casos de acção auto-derrotadora dos quais o do encontro de probabilidades que eu aqui apresentei é um exemplo. Estes são casos nos quais a acção sendo, de facto, auto-derrotadora, não é tal que seja de todo possível encontrar neles qualquer efeito de 'sucumbir à tentação' da gratificação iminente nem, portanto, pode neles estar plausivelmente em causa qualquer problema relacionado com o desconto das utilidades futuras ao seu valor presente. Na realidade, estes são casos de irracionalidade sincrónica e não de irracionalidade diacrónica. E são também casos nos quais o problema subjacente é claramente de natureza cognitiva e não motivacional. Simultaneamente, porém, o problema cognitivo que neles se manifesta não admite ser plausivelmente reconduzido a um diagnóstico de falha computacional circunstancial, seja ela simples ou complexa. Como poderemos então dar conta dele?

A minha resposta tentativa a esta questão é, resumidamente, a seguinte. As heurísticas rápidas e frugais que tendemos a usar com mais frequência, e que o programa de investigação de Gigerenzer tem vindo a trazer à luz do dia de forma sistemática, parecem ser implementadas sem, ou quase sem, controlo consciente. Quer dizer, elas parecem desembocar na produção de juízos rápidos, os quais determinam a nossa acção, independentemente de o que quer que seja que esteja também a ocorrer em simultâneo nas nossas mentes conscientes. Em geral, as soluções implementadas por estas heurísticas obtêm um enquadramento excelente nos ambientes nos quais elas emergiram. Mas o modo como estas soluções admitem ser generalizadas a outros ambientes depende fortemente da estrutura informacional subjacente à sua capacidade para alcançar sucesso adaptativo no ambiente original. Uma consequência desta dependência é, precisamente, o facto de que uma heurística bem adaptada à estrutura informacional que caracterizava um ambiente particular que foi típico no passado pode falhar quando esta estrutura muda subitamente de um modo essencialmente imprevisível num novo ambiente. E eu creio que é precisamente um falhanço deste género que presenciamos quando consideramos a resposta auto-derrotadora dos estudantes no primeiro cenário do experimento do encontro de probabilidades apresentado acima.

Partindo do princípio que esta explicação é plausível para este caso, será que podemos todavia generalizá-la a outros casos de

acção-derrotadora? O meu palpite é que sim, podemos. Deste modo, defendo que o tipo particular de abordagem cognitivista que aqui descrevo tem a capacidade de nos providenciar com um mecanismo que pode dar plausivelmente conta do modo como a acção auto-derrotadora é não apenas possível entre os humanos como, na realidade, deve ocorrer com frequência. A minha tese de carácter mais geral é, então, a de que a acção auto-derrotadora não-patológica é, pelo menos num grande número de casos, o resultado do disparo de uma heurística pertencente à nossa "caixa de ferramentas adaptativa" num ambiente que, não sendo o apropriado, é, não obstante, suficientemente semelhante ao ambiente apropriado para ter dado origem à mobilização pelo nosso aparelho cognitivo da heurística cognitivamente associada a este último. Defendo, ainda, que este mecanismo mental é suficientemente poderoso para, com frequência, levar a que a heurística cuja implementação é assim desencadeada suplante processos conscientes seguidos por nós para nos levarem a agir de modo diferente, mesmo em casos nos quais as decisões conscientes a que eles nos teriam conduzido poderiam ter sido mais apropriadas para lidar com o problema particular posto pelo novo ambiente.

Na realidade, esta dimensão do problema encontra-se igualmente presente no caso dos estudantes de Yale e do rato relatado acima. Com efeito, muitos dos estudantes que participaram na experiência tinham sido anteriormente sujeitos a um ensino explícito de princípios básicos de teoria da decisão. E, pelo menos alguns deles, relataram ter sentido um sentimento de surpresa quando tomaram consciência da resposta que deram, assim que os experimentadores os confrontaram com a discrepância entre os seus resultados e os do rato. A presença de um tal sentimento, associado à convicção íntima de que eles não conseguiam encontrar qualquer razão que justificasse o facto de não terem deixado prevalecer a resposta que decorreria logicamente do que tinham aprendido e que eles próprios consideravam a mais adequada, constitui precisamente o sinal fenomenológico típico da fraqueza de vontade. Creio, por isso, que faz todo o sentido considerar como bons exemplos de manifestação de fraqueza de vontade, no sentido do termo introduzido no início deste ensaio, casos de comportamento humano de encontro de probabilidades, que ocorram em circunstâncias nas quais a exibição desse comportamento se afasta de um modo

significativo da resposta normativa, e nas quais os seus protagonistas são agentes que aceitam de boa vontade a resposta normativa como sendo a sua melhor escolha nas circunstâncias e não são capazes de encontrar sinceramente um modo de justificar a sua escolha pela opção menos boa nessas circunstâncias.

A pretensão de generalidade da minha tese implica que o princípio explicativo que defendo deve também ser capaz de dar conta dos casos de desconto hiperbólico, isto é, dos casos em que são violadas as regras de consistência que permitem a racionalidade inter-temporal. E eu creio que esse é, de facto, o caso. Com efeito, e como o próprio Ainslie é levado a reconhecer, a maioria dos fenómenos de acção humana irracional cobertos pela sua explicação do desconto hiperbólico admite perfeitamente ser vista como a expressão actual de adaptações cognitivas que tiveram sucesso em ambientes que cessaram de ser predominantes nas sociedades humanas complexas, mas que tão-pouco desapareceram por completo. Só para vos dar um exemplo óbvio, o mecanismo de sucumbir à tentação certamente que foi com frequência adaptativo (e continua a sê-lo) na tarefa de tornar possível que a reprodução dos indivíduos humanos decorra com sucesso, uma tarefa cuja importância evolucionária julgo não carecer de demonstração. Se eu tenho razão neste aspecto, então parece-me ter ficado demonstrado que, enquanto que a abordagem proposta por Ainslie não tem a generalidade suficiente para dar conta de casos como o que eu introduzi, a minha proposta, para além de dar conta dos casos que eu introduzi, tem também um grau de generalidade suficiente para dar conta dos casos analisados por Ainslie. E creio que esta é uma vantagem não negligenciável da minha proposta.

5. Racionalidade divina, racionalidade de segunda divisão e racionalidade situada

Gostaria ainda de acrescentar, como observação final, que os conceitos de racionalidade indirecta de Elster e Ainslie parecem-me ter ainda sido definidos sob a influência de uma visão de "cima para baixo" do Homem, a origem da qual é, em última análise, de carácter teológico. De acordo com esta visão, o Homem seria uma criatura deixada a meio caminho entre os reinos divino e animal e, em virtude desse facto, participando dos dois. O facto de ter sido

feito à imagem do seu criador divino permitir-lhe-ia ser racional; a sua mortalidade e as suas limitações intrínsecas impedi-lo-iam de ser inteiramente racional. Ele teria assim que se desenvencilhar com uma forma própria de racionalidade imperfeita ou de segunda divisão. Elster e Ainslie tentaram capturar esta forma peculiar de racionalidade no âmbito dos seus esforços teóricos.

Creio que chegou a altura de substituir um tal conceito de racionalidade imperfeita, ou de segunda divisão, por um conceito evolucionário de racionalidade, definido de "baixo para cima". Os conceitos de racionalidade situada, primeiro introduzido por Herbert Simon, e de racionalidade ecológica, avançado mais recentemente por Gigerenzer e o seu grupo, parecem-me constituir passos seguros nessa direcção.

Como vimos, os proponentes destes conceitos de racionalidade defendem a tese de que o uso de procedimentos heurísticos pelos nossos sistemas cognitivos é inevitável. Mas o que eu creio ser especialmente distintivo na sua abordagem é, não propriamente a defesa desta tese, a qual eu creio ter-se já tornado praticamente indisputável, mas, antes, a tese, substancialmente mais forte, que eles também defendem, de que um tal uso não resulta simplesmente do facto de os nossos aparelhos cognitivos terem sido obrigados a desenrascar atalhos mais rudes e menos precisos que os caminhos racionais ideais para conseguirem compensar a impossibilidade física de acederem a recursos ilimitados. De facto, de acordo com o ponto de vista ao qual eles pretendem dar expressão, as heurísticas rápidas e frugais, quando empregues no âmbito das estruturas ambientais apropriadas, e debaixo dos constrangimentos a que um sistema cognitivo de carácter biológico se encontra, nelas, sujeito, são, frequentemente, mesmo quando consideradas em termos absolutos, modos mais robustos de lidar com um mundo incerto e complexo do que as ferramentas de optimização tradicionais. Ora, é precisamente *esta* contenção que, creio, muda substancialmente a natureza do debate filosófico em torno da compreensão da acção humana.

Referências

Ainslie, G. 1992. *Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person.* Cambridge: Cambridge University Press.

Ainslie, G. 2001. *Breakdown of Will.* Cambridge: Cambridge University Press.

Davidson, D. 1970. 'How is Weakness of the Will Possible?' in Feinberg, J. (ed.), *Moral Concepts.* Oxford: Oxford University Press. Reprinted in Davidson, D. 1980, *Essays on Actions and Events.* Oxford: Clarendon Press, 21-42.

Davidson, D. 1982. "Paradoxes of Irrationality" in Wohlheim, R. & Hopkins, J. (eds.): *Philosophical Essays on Freud.* Cambridge: Cambridge University Press.

Elster, J. 1984. *Ulysses and the Sirens – Studies in Rationality and Irrationality.* Cambridge: Cambridge University Press.

Elster, J. 2007. *Explaining Social Behavior – More Nuts and Bolts for the Social Sciences.* Cambridge: Cambridge University Press.

Gallistel, C.R. 1990. *The Organization of Learning.* Cambridge (MA): The MIT Press.

Gigerenzer, G. 2000. *Adaptive Thinking – Rationality in the Real World.* Oxford: Oxford University Press.

Gigerenzer, G. 2004. 'Fast and Frugal Heuristics: the Tools of Bounded Rationality'. In Koehler & Harvey (eds.), *Handbook of Judgment and Decision Making.* Oxford: Blackwell.

Gigerenzer, G., Todd, P.M., & the ABC Research Group 1999. *Simple Heuristics that Make Us Smart.* Oxford: Oxford University Press.

Simon, H. 1957: *Models of Man - Social and Rational.* New York: Wiley & Sons.

Simon, H. 1983. *Reason in Human Affairs.* Stanford (CA): Stanford University Press.

Wilson, T.D. 2002: *Strangers to Ourselves – Discovering the Adaptive Unconscious.* Cambridge (MA): The Belknap Press of the Harvard University Press.

Zilhão, A. 2006. "Incontinence, Fast and Frugal Heuristics and Probability Matching" in Manrique, F.M. y Peris-Viñé, L.M. (eds.): *Actas del V Congreso de la Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España.* Granada: Ediciones Sider, 242-246.

Zilhão, A. 2009. "Incontinence, Honouring Sunk Costs, and Rationality" in Suarez, Dorato, Redei (eds.), *EPSA Philosophical Issues in the Sciences.* Berlin: Springer Verlag, 303-310.

Zilhão, A. 2010. *Animal Racional ou Bípede Implume? – Um Ensaio sobre Acção, Explicação e Racionalidade*. Lisboa: Guerra & Paz.

Zilhão, A. 2011. "Acção, Decisão e Explicação da Acção" in Cadilha, S. & Miguéns, S. (coord.): *Acção e Ética – Conversas sobre Racionalidade Prática.* Lisboa: Colibri, 97-142.